OPENFABRICS ALLIANCE

Exploring Improvement to Verbs

Tom Stachura, OFA TAC Co-chair

10TH ANNUAL INTERNATIONAL DEVELOPER WORKSHOP

#OFADevWorkshop

About this session…

- I am not a Verbs expert
- But that's ok, b/c **YOU** (& Nathan) are…

- I co-chair the TAC (Technical Advisory Council)
- The TAC is responsible for driving OFA direction

- I need your help to get through this session
- Audience participation is required

Commercial Spot: TAC

- TAC = Technical Advisory Council
- TAC Charter, the fine print:
  - Investigate technology trends
  - Review needs of end user markets/apps/technology
  - Maintain links to IBTA TWG, Spec bodies, & end users
- TAC Charter, my words:
  - ***Find OFA growth vectors***
- We (a group of smart technologists + me) meet twice a month
- I drive the agenda

TAC Outputs & Focus Areas

- Key TAC Outputs to-date
  - Vetted new ULPs & Proposing OFA synergies
  - Analyzed the hi-performance needs of Applications
  - OFI WG was incubated in the TAC
- Current TAC Exploration Areas:
  - Improving Verbs
  - Expanding to the Cloud
  - Storage Usage models, esp. NVM

Contact tom.l.stachura@intel.com if you are interested

End of Commercial

- Key TAC Outputs to-date
  - Vetted new ULPs & Proposing OFA synergies
  - Analyzed the hi-performance needs of Applications
  - OFI WG was incubated in the TAC
- Current TAC Exploration Areas:
  - **Improving Verbs**
  - Expanding to the Cloud
  - Storage Usage models, esp. NVM

The Hurdles with Verbs…

- Nathan Hjelm @ LANL provided great feedback:
  1. RDMA-CM doesn't scale
     - Issues scaling beyond 1500 ranks and 32 CPUs/node. SSA?
  2. RC mode runs out of queue pair resources
     - Good focus here (DCT), but no standardization
  3. Verbs interfaces don't map well to MPI semantics
     - Supporting multiple MPIs causes code bloat
  4. Heavy cost of setup & managing memory registration
     - More of an issue for PGAS
  5. Lack of standardization between h/w implementations
     - i.e. PSM vs. MXM
  6. No mapping to "Well-known ports"
     - QPn is random – MPI w/ UD wants a specific port & QPn

Verbs Hurdles – A Simplification

- Scalability
  - RDMA-CM
  - QP Resources
- Scalability & Usability
  - Heavy cost for memory registration
- Application Impedance Mismatch
  - Not mapping to well to MPI
  - Different h/w implementations
  - Mapping to “well-known” ports

Any key hurdles we missed?

Verbs Hurdles – We heard it this week…

- Scalability
  - RDMA-CM
  - QP Resources

NASA Pleades, DCT, OFI WG, MPI API, SSA

- Scalability & Usability
  - Heavy cost for memory registration

OFI WG, MPI API PGAS API, ODP

- Application Impedance Mismatch
  - Not mapping to well to MPI
  - Different h/w implementations
  - Mapping to “well-known” ports

OFI WG
MPI API
SMC-R

Verbs Hurdles – Next Steps…

- Scalability
  - RDMA-CM
  - QP Resources
- Scalability & Usability
  - Heavy cost for memory registration

Recommended Focus area for OFA & IBTA

- Application Impedance Match
  - Not mapping to well to MPI
  - Different h/w implementations
  - Mapping to “well-known” ports

OFI WG Focus Area

My Only Presentation Picture

Thank You

OPENFABRICS ALLIANCE

10TH ANNUAL INTERNATIONAL DEVELOPER WORKSHOP